

Resource Related Research - Computers & Chemistry

ANNUAL REPORT

August 1, 1976 - April 30, 1977

Stanford University  
NIH/BRP Grant RR-00612

Carl Djerassi, Principal Investigator  
(Social Security No. [REDACTED])

1 OVERVIEW OF RESEARCH ACTIVITIES

The past year's activities in computer applications to chemical problems have continued the progression of new research, followed by applications and export to a wider community of scientists. The simplest way to detail this work is to place it within the framework of the larger problem of elucidation of unknown molecular structures. Our research, development and future plans focus on both the question of structure elucidation in general and the problem of providing computer assistance to scientists engaged in specific aspects of this important activity.

A simplified representation of major milestones in solving unknown biomolecular structures by manual methods is presented in Figure 1.

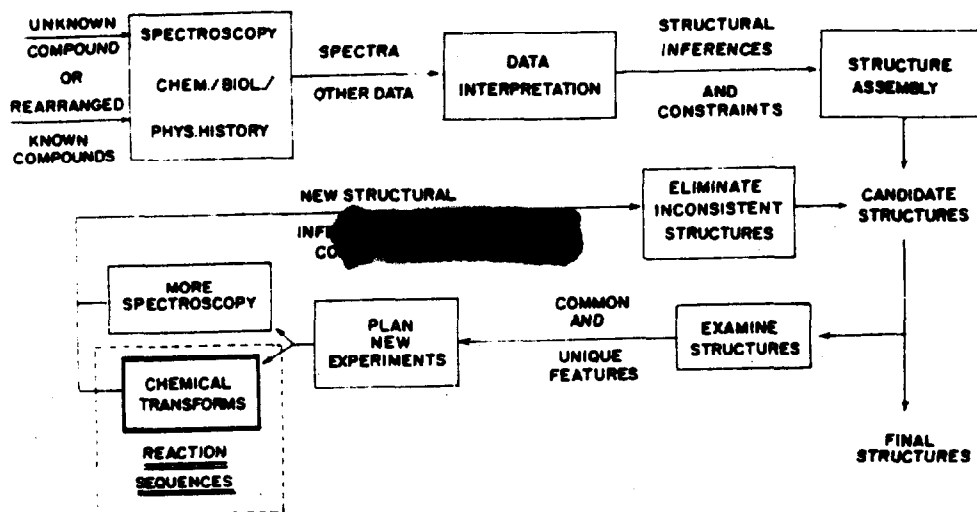


Figure 1. Important steps in manual solution of structures of unknown chemical compounds.

These steps, indicated as separate boxes, may be performed explicitly or implicitly. There are considerably more complex relationships among the boxes of Fig. 1 than are indicated when structures are actually solved. Nevertheless, the Figure provides a good introduction to both our recent work and our future directions. We describe briefly each of the milestones in the following paragraphs. More detailed discussions of each topic follow in subsequent sections.

The first step in identification of an unknown structure is to separate it from other components in a potentially complex mixture and to isolate it in reasonably pure form. These steps are performed by scientists, frequently with the assistance of various instruments. Although our research is not directed toward any part of this separation and isolation procedure (except insofar as these procedures also yield data which are subject to computer-assisted interpretation), information about the chemical and physical characteristics of the compound may be crucial to further efforts to determine its structure.

Depending on the quantity of sample available and its characteristics, various spectroscopic and additional chemical data are then collected on the unknown. A mass spectrum is frequently obtained, e.g., from a combined gas chromatograph/mass spectrometer (GC/MS) system. An important part of our recent proposal to the NIH is directed toward automation of combined GC/MS systems operated at high mass spectrometer resolving powers. Data on elemental compositions and relative

ion abundances are then available in computer-readable form for further analysis (see MSRANK). The chemist possess an armamentarium of spectroscopic techniques which can be brought to bear on a structure. One advantage of our work is that any data so obtained can be used to help solve the structure as long as it can be expressed, manually or by computer, in substructural statements about the unknown.

The next important phase in structure elucidation is interpretation of the available data (Fig. 1) in terms of structural features of the molecule. These interpretations may be in terms of known structural units ("superatoms", polyatomic aggregates of atoms in known configurations), or in terms of structural units, ring sizes, proton or carbon distributions. The latter set of features represents constraints on the kinds of structures which are possible. Our efforts in the area of computer-assisted data interpretation are focussed on mass spectral and carbon-13 nuclear magnetic resonance (<sup>13</sup>CMR) data. We are developing general approaches to automated analysis of these data in terms of structural features of unknowns.

Our recent efforts are summarized in Figure 2, and discussed in detail subsequently. We have been concerned with use of these data from two points of view, planning and prediction (Fig. 2). During planning, experimental data are examined in order to extract specific structural information to be used in assembling candidate structures. In prediction each candidate structure is tested to determine how closely its predicted spectrum agrees with the observed spectrum. The candidates can be ranked accordingly. The Meta-DENDRAL research is directed toward determination of rules of spectroscopic data which can be used either for planning or prediction (see below).

DATA INTERPRETATION"PLANNING"

EXTRACTION OF STRUCTURAL  
INFORMATION DIRECTLY FROM  
SPECTROSCOPIC DATA.

1. MASS SPECTRA - MDGGEN
2. <sup>13</sup>CNMR

PREDICTION

USE OF SPECTROSCOPIC  
DATA TO RANK  
CANDIDATE STRUCTURES.

1. MSPRUNE, MSPRED
2. <sup>13</sup>CNMR

↙ ↘  
META - DENDRAL

FORMATION OF RULES TO BE  
USED FOR BOTH PLANNING  
AND PREDICTION.

Figure 2. Relationship between use of rules in either planning or prediction. Both approaches are used in utilizing data for structure elucidation.

Given possible structural fragments of the complete molecule and constraints on how these fragments may be assembled into complete molecules, a process of structural assembly follows (Fig. 1). There has been no proven algorithm for solving this problem prior to earlier work supported by the current grant. Traditionally, this process has been left to manual, pencil and paper work. Our CONGEN program, which was designed to solve this problem, is farthest advanced of programs designed to assist in various aspects of structure elucidation. It performs the structural assembly process, under constraints, and allows the scientist using the program to examine structural candidates and remove those deemed implausible (Fig. 1). A large portion of our recent and future work is directed toward improving the CONGEN program and building other facilities around it (see later sections). We have demonstrated the utility of CONGEN in structural studies, and subsequent sections discuss our recent developments and applications of CONGEN as well as our interactions with other scientists desiring access to our programs.

Given a set of structural candidates, the experimenter examines them to determine what experiments might be performed to focus on the correct structure by stepwise rejection of alternative hypotheses. When there are only a small number of possibilities under consideration, manual methods suffice. But CONGEN provides the capability for exhaustive enumeration of structural possibilities at a point in a structural problem when there may be many hundreds of possibilities. It is very difficult to examine these structures and plan experiments by hand. We have begun exploring ways to provide computer assistance to this important aspect of structure elucidation. We refer to this research area as the Experiment Planner, discussed in more detail below.

When new experiments have been planned the researcher carries them out and uses the results as additional constraints on the structural candidates (Fig. 1). New experiments may include collecting of additional spectroscopic data or performing a sequence of chemical reactions on the unknown. The latter experiments may be chosen to convert the unknown into a related compound which possesses physical or chemical properties more amenable to analysis. During the past year we have developed a program to assist scientists in carrying out representations of chemical reactions in the computer and eliminating undesired structural candidates based on constraints exercised on the products of the reaction. This work is described in two subsequent sections. One section describes use of the program, which we call REACT, to explore structural possibilities exactly as outlined above. A later section describes recent progress in increasing the power of REACT.

## 2 EXPERIMENT PLANNER

We have begun preliminary considerations of design and implementation of an experiment planner. This program will assist chemists in designing the most effective set of experiments to perform to solve the structure. Although the experiment planner will be a future activity of our group, we are developing and using other structure manipulation functions which will provide groundwork for future developments.

One important aspect of experiment planning is the ability to examine in some way the set of candidate structures. Although

many can be drawn for visual review, drawing is impractical when dozens or hundreds of structures are involved. To assist persons using CONGEN in reviewing their structures we have developed a function auxiliary to CONGEN which we call SURVEY.

## SURVEY

FUNCTION: AID IN PERCEPTION OF ANY OF A  
PRE-SPECIFIED SET OF STRUCTURAL  
FEATURES IN A GROUP OF  
STRUCTURAL CANDIDATES.

E.G. A) FUNCTIONAL GROUPS  
B) TERPENOID SKELETONS  
C) AMINO ACID SKELETONS

Figure 3. Function of the SURVEY program and examples of recent application areas.

The function of SURVEY is summarized in Figure 3. SURVEY simply acts as a reminder to the scientist of the presence or absence of certain structures or structural features. During the past year we have used SURVEY extensively. For example, we have used it to detect implausible functional groups in a set of candidate structures, using a file of substructures representing a wide variety of functionalities. In many problems, implausible functional groups are forgotten and CONGEN is never constrained to remove them. Another example of use of SURVEY is in conjunction with collaborative work with persons in the Department of Genetics. In analysis of serum or urinary metabolites in patients of high risk of metabolic disorder, we have had occasion to use CONGEN in exploration of unknown structures [Report HPP-77-11]. Some of these structures could formally be conjugates of amino acids with organic acids. If so, such structures will possess backbones of naturally-occurring amino acids. SURVEY was used to provide a summary of which structural candidates possessed such amino acid skeletons.

We have recently used SURVEY in a related application involving the structure of "polyalthenol", discussed by LeBoeuf,

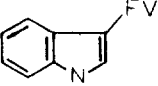
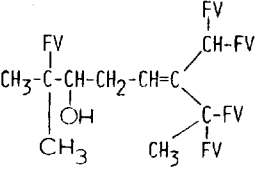

et al. (Figure 4). Superatoms and constraints supplied to CONGEN to derive structural candidates are summarized in Fig. 4.

We summarize in Figure 5 the structural possibilities which resulted. There are five structures possessing a bicyclo[2.1.1] system, and six which possess a bicyclo[4.3.1] system (Fig. 5, top). These structures are energetically less favorable. For example, several possess a double bond at a bridgehead atom, which violates Bredt's Rule. There remain, however, 11 structures which are not formally excluded by data presented by LeBoeuf, et al. Because these workers based their structural assignment on biogenetic grounds, we used SURVEY and REACT to test their hypothesis. We have, in computer-accessible libraries, known terpenoid ring systems which can be used within SURVEY to test sets of structures for known skeletons. None of the 22 structural candidates possesses a previously known skeleton. Because the authors postulated a relationship to a known skeleton via a single methyl shift, we used REACT to exercise a single methyl shift in all possible ways on each of the 22 candidates. SURVEY was then used to test the results for the presence of known terpenoid systems, and the drimane skeleton, the postulated precursor of polyathenol, was the only known skeleton which resulted. This does not prove the hypothesis of LeBoeuf, et al., but certainly helps strengthen it.

SURVEY is, however, only the barest beginning of an experiment planner, even though it has proven useful. We plan to build from this beginning toward a much more powerful system.

M. LeBOEUF, M. HAMONNIÈRE, A. CAVÉ, H. GOTTLIEB, N. KUNESCH,  
AND E. WENKERT, TET. LETT., 3559 (1976).

"POLYALTHENOL"  $C_{23}H_{31}NO$

SUPERATOMS	ARBITRARY NAME	NUMBER
	IN	1
	BI	1
CH <sub>3</sub> -FV	ME	1
FV-CH <sub>2</sub> -FV	CH <sub>2</sub>	3
	CH	1

#### CONSTRAINTS

- 1) ALL FREE VALENCES BONDED TO NON-HYDROGEN ATOMS
- 2) GOODLIST
 

	IN-CH <sub>2</sub> -BI	1 TO ANY
(EVENTUALLY	IN-CH <sub>2</sub> -CH <sub>0</sub> →0)	
	ME-(BI CH)	1 TO ANY
(EVENTUALLY	CH <sub>3</sub> -CH, EXACTLY 1)	
- 3) GOODRINGS
 

2	EXACTLY 5
---	-----------
- 4) BADRINGS
 

3	
---	--

Figure 4. Superatoms and constraints supplied to CONGEN in investigations of plausible structural alternatives to the proposed structure of Polyalthenol.



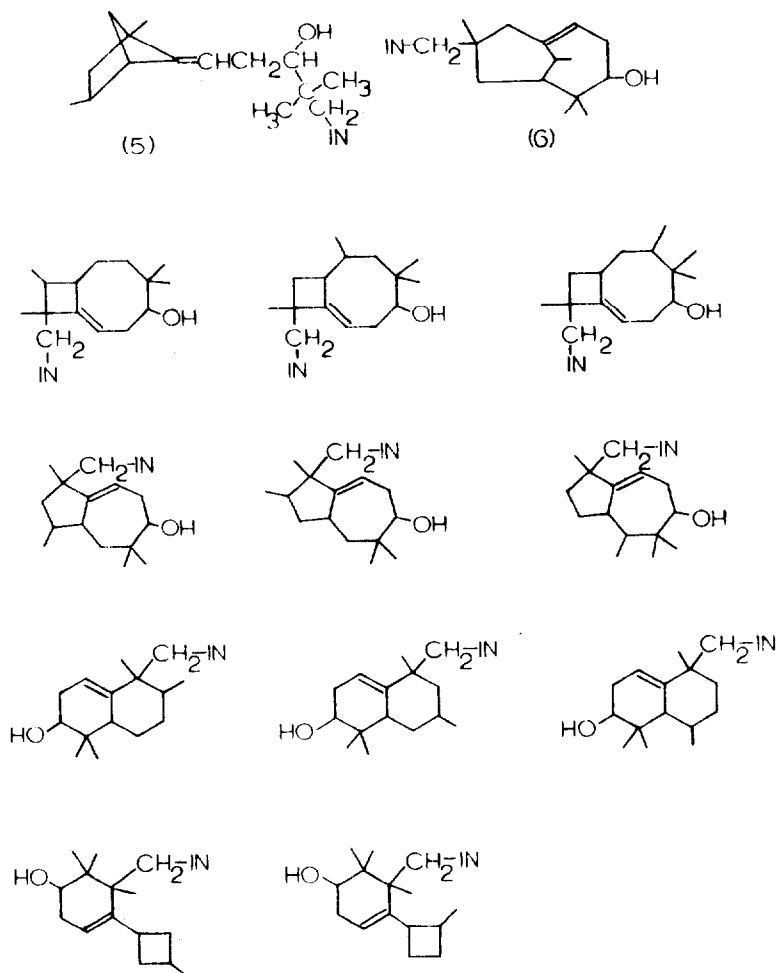


Figure 5. Structural candidates for Polyalthanol based on data given in Figure 4.

### 3 Applications of REACT to Structure Elucidation Problems

We have recently described our initial efforts toward representation of chemical reactions and their use in structure elucidation problems [Report HPP-76-5]. These efforts provided the framework for carrying out reactions within the computer which emulate actual laboratory reactions performed on a unknown. Constraints on the numbers and identities of the products are used to constrain the reaction products and, implicitly, the starting materials. Based on the results of that work we drew up a set of steps to be carried out to provide a truly useful tool for the chemist. Although the current program can be used in

applications to real problems it has some fundamental limitations which we have been working to solve. The developments we have undertaken to improve REACT are summarized in Figure 6.

#### REACTION CHEMISTRY DEVELOPMENTS

1. SEPARATION FROM CONGEN - COMMUNICATION VIA FILES OF STRUCTURES.
2. ADDING CONSTRAINTS - SITE - AND TRANSFORM - SPECIFIC.
3. CONTROL STRUCTURE - RAMIFICATION
  - A. ESTABLISH RELATIONSHIPS AMONG PRODUCTS AND REACTANTS
  - B. DEAL PROPERLY WITH RANGES OF NUMBERS OF PRODUCTS
4. INTERACTION - DEVELOP MANIPULATION COMMANDS WHICH
  - PARALLEL LABORATORY OPERATIONS, E.G.,
  - SEPARATE INTO FLASKS, TEST CONTENTS OF
  - VARIOUS FLASKS, INCOMPLETE SEPARATIONS,
  - ETC.
5. REPRESENTATION OF REACTIONS
6. PROSPECTIVE DETECTION OF DUPLICATE PRODUCTS BASED ON SYMMETRY PROPERTIES OF: A) STARTING MATERIAL; AND B) TRANSFORMATION.

Figure 6. Current and future direction for improvement and extension of REACT, a program for exploration of applications of reaction chemistry to structure elucidation problems.

We first undertook to separate REACT from CONGEN, for two reasons. One reason was due to program size. Many functions of CONGEN are not needed in REACT and become unnecessary when only REACT is being exercised. The procedures of structure generation (CONGEN) and REACT are sequential and a separate program introduces no problems. A second reason was the different uses of certain CONGEN functions in REACT. For example, the ways in which the graph matcher is used are different between the two programs, necessitating keeping two different versions around with the programs together. The separation has been accomplished. The current version of REACT is now a separate program. It communicates structural information with CONGEN via files. All interactive portions are consistent with the structural manipulation functions of CONGEN so that learning the structural language of CONGEN is sufficient to use either program.

We have also added new constraint types to the reaction to expand greatly the ways in which reactions can be defined and constrained. An example of new extensions to reaction definitions illustrates some of the new features (Figures 7-10). The reaction defined here is one which will perform a dehydration of an alcohol; the site of the reaction is defined in Fig. 7.

```
:EDITREACT  
NAME: DEHYDRATION  
(NEW REACTION)
```

```
*SITE  
>CHAIN 3  
>ATNAME 1 0  
>HRANGE 1 1 1 3 1 3  
>ADRAW
```

DEHYDRATION: (HRANGES NOT INDICATED)

O-C-C

>DONE

```
*TRANSFORM  
>UNJOIN 1 2  
>JOIN 2 3  
>DELATS 1  
>ADRAW
```

DEHYDRATION: (HRANGES NOT INDICATED)

C=C

>DONE

Figure 7. Definition of reaction site and chemical transform in REACT.

The transform is defined as cleavage and loss of the oxygen resulting in formation of a double bond between the two carbon atoms of the original site (Fig. 7). In this particular dehydration the chemist wished to specify a site-specific constraint. It was known that a tertiary butyl group was part of the structure, and the dehydration will be prevented if that group is in close proximity to the reaction site (i.e., in a position alpha to the carbinol carbon).

```

*DEFINE-CONSTRAINTS
: ?
PLEASE ENTER ONE OF:
GRIPE          BUGOUT          GENERAL(G)    SITESPECIFIC(S)
TRANSFORMSPECIFIC(T)        DONE          HALT

: SITESPECIFIC
NAME: HINDERED
(NEW CONSTRAINT)
(WARNING: THE FINAL CONSTRAINTS MUST HAVE AT LEAST ONE ATOM OF THE
SITE)
> NDRAW

HINDERED: (HRANGES NOT INDICATED)
NON-C ATOMS: 1 0

1-2-3

> BRANCH 3 2 4 1 4 1
> ADRAW

HINDERED: (HRANGES NOT INDICATED)

      C
      |
O-C-C-C-C
      |
      C

> DONE

```

Figure 8. Definition of a site-specific constraint to be applied to the reaction DEHYDRATION.

The definition of this constraint is given in Figure 8. Subsequently, this constraint ("HINDERED") is placed on BADLIST for constraints specific to the site as shown in Fig. 9. The completed definition of the reaction is summarized in Figure 10.

\*CONSTRAINTS

:?

PLEASE ENTER ONE OF:

GRIPE

BUGOUT

ST FOR CONSTRAINTS ON STARTING MATERIAL

S FOR SITESPECIFIC CONSTRAINTS

T FOR TRANSFORMSPECIFIC CONSTRAINTS

PR FOR CONSTRAINTS ON PRODUCTS

DONE

HALT

:S

>BADLIST

BADLIST CONSTRAINTS

CONSTRAINT NAME:HINDERED

CONSTRAINT NAME:

-----

>DONE:DONE

Figure 9. Specification of constraint named HINDERED as a BADLIST constraint for the reaction.

```

*SHOW
SITE:
NAME=DEHYDRATION
ATOM# TYPE ARTYPE NEIGHBORS HRANGE
  1   O  NON-AR   2         1-1
  2   C  NON-AR   1 3         1-2
  3   C  NON-AR   2         1-2

DEHYDRATION: (HRANGES NOT INDICATED)
NON-C ATOMS: 1 0

1-2-3

TRANSFORM:
  UNJOIN 1 2
  JOIN 2 3
  DELATS 1

DEHYDRATION: (HRANGES NOT INDICATED)

2=3

CONSTRAINTS:
CONSTRAINTS ON STARTING MATERIAL:
NO CONSTRAINTS
SITE-SPECIFIC CONSTRAINTS:
-----
BADLIST CONSTRAINTS
  NAME
  HINDERED
-----
TRANSFORM-SPECIFIC CONSTRAINTS:
NO CONSTRAINTS
CONSTRAINTS ON PRODUCTS:

NO CONSTRAINTS
*DONE
(DEHYDRATION DEFINED)
(DEHYDRATION ADDED TO THE REACTION LIST)

```

Figure 10. Summary of the completed definition of the DEHYDRATION reaction.

The remaining items summarized in Figure 6 are currently under development. We are redesigning the control structure so that the scientist using the program can use intuitive concepts as commands, such as separation. To carry this out important parts of the current mechanism have to be redesigned. Although the current program can be used effectively, its non-intuitive approach to dealing with reactions yielding multiple products and subsequent separation (within the computer) and analysis of each product presents a barrier to use by a wider community. We are continuing to develop our capabilities for representing reactions to ensure that the user of REACT has a complete descriptive language with which to specify reactions. We continue to study ways to avoid duplication in carrying out reactions. We know how to implement certain of the symmetry-related constraints and will do so shortly.

#### 4 CONGEN Developments

The problem solving paradigm that has emerged from DENDRAL work is the so-called "plan-generate-test" paradigm. It is based on heuristic search of a space of possible hypotheses with planning before generation of hypotheses and testing of each generated candidate.

The generator for DENDRAL, named CONGEN, is a general-purpose graph generator which produces a list of all possible graphs containing specified numbers of nodes of various types. The most important features of the generator are that the list of graphs is guaranteed to be complete and non-redundant and, equally important, that the list need not be exhaustively generated. The generator can be constrained to produce only graphs that meet specified criteria that are inferred from the initial problem data.

During the past year, CONGEN has developed along two major lines: 1) tools have been developed which will allow more efficient and "intelligent" use of substructural information supplied by the chemist; and 2) data from chemical reactions and from observed mass spectra can be used to eliminate unlikely structural candidates from a set produced by a CONGEN generation. These extensions will be discussed below.

##### 4.1 Intelligent use of constraining substructural information

There is sometimes a significant conceptual gap between the intuitive chemical phrasing of a CONGEN problem and the phrasing which is most efficient, in both computer time and storage requirements, for the program. CONGEN provides a rich language for stating structure elucidation problems in precise substructural terms. However, there are usually many ways of defining a given problem and different definitions can place widely different demands upon the program. We have a continuing interest in reducing this conceptual gap by in making CONGEN responsible for rephrasing a problem in the most efficient way, thus freeing the chemist to concentrate upon the chemical, rather than the algorithmic, aspects of a given case.

One distinction which is frequently puzzling to new CONGEN users is the one between superatoms and GOODLIST items. A superatom is a polyatomic "building block" which CONGEN joins with other superatoms and single atoms to form full structures. GOODLIST items are substructures which are required to be present in those full structures, but they are not incorporated directly into the initial phrasing of a problem as are superatoms. Rather, their presence or absence is tested by a graph-matching

routine after the structures are produced. Frequently, a great many structures produced by the structure generator are discarded by this final test and a significant amount of the program's time can be spent "shooting blanks". The concepts behind these two types of constraints - that specified substructural features must be present - are similar, but their implementations differ substantially in efficiency.

GOODLIST items cannot simply be transferred to the superatom list, though, because GOODLIST items are allowed to share atoms and bonds with other GOODLIST items or with superatoms. For example, if two substructures which are benzene rings are placed on GOODLIST, then a naphthalene derivative will be an acceptable structure even though the two occurrences of the ring have two atoms and one aromatic bond in common. Because of the building-block nature of superatoms, they may be joined to one another by additional bonds in CONGEN, but never "merged" (i.e., overlapped). Thus the price of efficiency is a more restricted interpretation of structural possibilities for superatoms.

We have developed a new procedure which captures the best of both situations. In order to incorporate a GOODLIST substructure into the problem at the earliest stage, it is necessary to find all unique ways that the given substructure can be created using parts of the existing building blocks (atoms and superatoms). This produces a set of new CONGEN problems with more or larger superatoms, each of which is easier to solve than the original one because the GOODLIST item is built-in and needs not be tested. Figure 11 shows schematically some of the ways this construction might occur: a) by bonding together two (or more) existing superatoms to create one larger one; b) by bonding additional atoms to a superatom to create a larger one; and c) by constructing a copy of the substructure from single atoms, creating a new superatom.



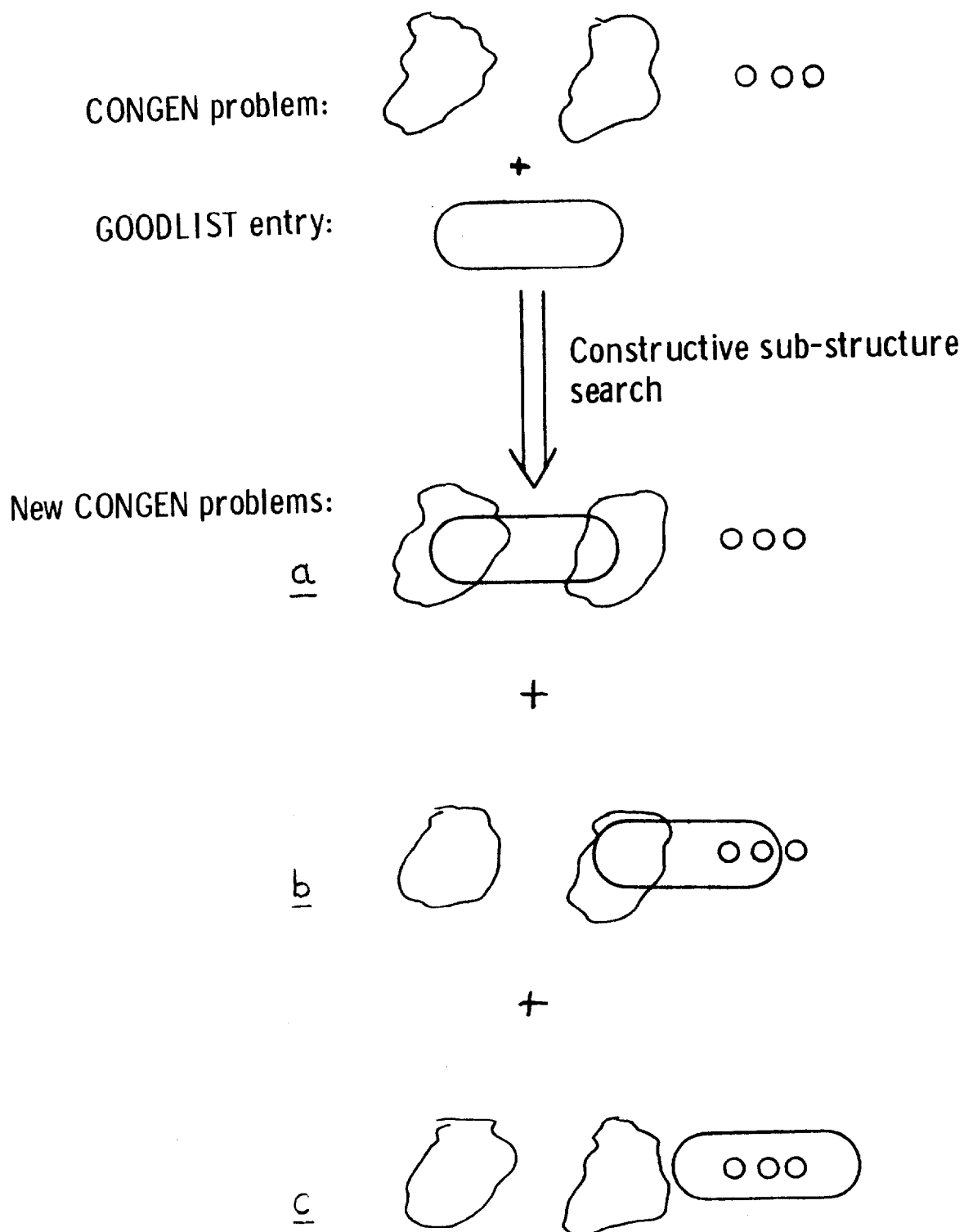
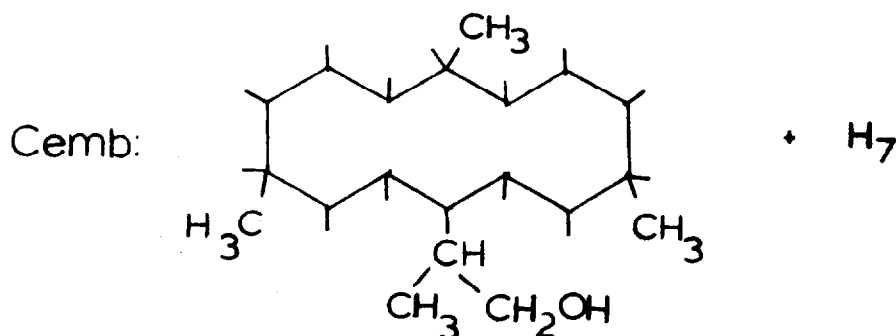


Figure 11. Example of breaking one GOODLIST substructure into several subproblems for CONGEN, each with different superatoms.

The algorithm is derived from the CONGEN graph-matching routine with the additional feature that as it searches for the substructure it is allowed to create new bonds (up to the limit of available new bonds in the original CONGEN problem) whenever they are necessary for the search to proceed. During the search, full account is taken of the topological symmetry of the superatoms in the original problem so that fittings which are redundant with respect to these symmetries are avoided. The substructure itself may possess some symmetry as well, but this is currently not considered.

Figure 12 summarizes a CONGEN problem which was attempted but which could not be completed because of the unintelligent use of GOODLIST. The problem amounts to finding all ways of allocating three new bonds to the free valences (the bonds with unspecified termini) in the superatom CEMB such that the three indicated substructures are present in the final molecules. There are perhaps 10,000 unique allocations of those three new bonds, but only 7 pass the GOODLIST tests. Using GOODLIST as a post-test only, CONGEN would generate all 10,000 and discard nearly all of them, a process which would have been so lengthy that it was never completed. The constructive graph-matching routine approaches the problem in a much more efficient and chemically intuitive way: 1) there are only three places in which the first GOODLIST item can be constructed; 2) for each of these, there are four ways of constructing the second; and 3) for each of these, there are 0, 1 or 2 ways of incorporating the third. It quickly arrives at the correct set of solutions.



GOODLIST:

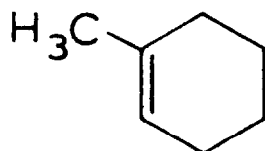
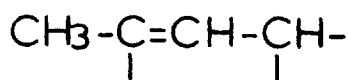
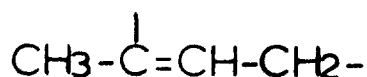


Figure 12. Example showing the inefficiency of specifying a constraint as a GOODLIST item instead of analyzing its implications for constructing allowable chemical graphs.

Most CONGEN problems contain one or more GOODLIST items which can be processed in this way, and when the constructive graph-matcher is fully integrated into CONGEN, it will make a substantial difference in its ability to use this structural information effectively.

#### 4.2 New tools for post-pruning CONGEN structures.

From an algorithmic standpoint, CONGEN is successful if it can, in a reasonable amount of time and without exhausting storage resources, produce a list of candidate structures satisfying the chemist's constraints. However, this list is often quite large, perhaps several hundred structures, and from a chemical standpoint the problem may be far from complete. It

remains for the chemist to discriminate among the candidates, eventually reducing the possibilities to just one structure. A SURVEY function is available for classifying the list into groups of chemically related structures using either pre-defined or user-defined libraries of substructural features, and this process can help the chemist perceive groups which might easily be ruled out by additional experiments. Also, the graph-matching (pruning) mechanism of CONGEN allows him to express, in terms of substructural tests on the candidates, new data which he gathers on the unknown. These are both important aids in dealing with a list of candidates, but are restricted to tests which can easily be phrased purely in terms of structural features of the candidates themselves.

There are two informative sources of data which cannot always be phrased in this way: 1) structural features observed in products of the unknown when it undergoes simple chemical reactions; and 2) empirical spectroscopic measurements on the unknown which cannot be interpreted unambiguously in precise structural terms. During the past year, we have made progress in utilizing such information. The program REACT addresses the first problem while MSRANK concerns the second, in the context of mass spectrometric observations.

#### 4.2.1 REACT

This program [see Report HPP-76-5] has two basic goals: 1) to provide the chemist with a computerized language for defining graph transformations and applying them to structures, thus simulating chemical reactions; and 2) to automatically keep track of the interrelationships between structures in a complex sequence of reactions so that whenever structural claims are made ruling out structures at one level, the implications in terms of structures at other levels can be traced. During the last year some progress has been made toward both of these goals.

EDITREACT, the reaction-editing language, has been extended to allow the user to define subgraph constraints which apply relative to a potential reaction site rather than to the molecule as a whole. For example, in the present version of REACT, we can say either that a hydroxyl group (OH), if present anywhere in the reactant molecule, would inhibit the reaction, or that such inhibition would take place only if the OH group is adjacent to the reaction site. Such site-specific constraints, applied either before or after the transformation (i.e., reaction) has been carried out on the site, are critical to the detailed description of real chemical reactions. The inclusion of this facility in REACT substantially increases its usefulness in real-world chemical problems.

The bookkeeping problem has undergone a complete

reconceptualization in the past year, the purpose being to mimic more closely the actual steps taken by a chemist in the laboratory. In the initial implementation, a set of products arising from the application of a given reaction to a given starting structure could be subjected to a multi-level classification which grouped the products based upon user-defined substructural constraints. Each of these classes had an associated minimum and maximum number, representing the numbers of products which were allowed to be members of the class. Any starting materials whose products could not satisfy these conditions were removed from the list of candidates. Structures in any class could be further reacted, their products classified, and so on. This treatment of bookkeeping was sufficient for stating many chemical problems. For example, suppose a chemist knew that a particular reaction on an unknown compound yielded two carbonyl compounds (i.e., containing C=O), at least one of which was an ester (-O-C=O). He could define a product class CARBONYL using the C=O substructure with a minimum and maximum of two products. He could then define a sub-class of CARBONYL called ESTERS using the substructure -O-C=O with a minimum of one and a maximum of two products. The program would automatically use this information to eliminate candidate starting structures which could not give the indicated product distribution with the given reaction.

There are chemical problems, though, for which the above scheme is too rigid. For example, suppose a reaction gives several products, two of which are isolated and labelled P1 and P2. Suppose that only a small amount of P1 is available so only mass spectroscopic measurements are practical. Suppose also that a deuterium-exchange experiment shows that P1 has two exchangeable protons (say, either N-H or O-H). P2 shows a strong carbonyl absorption in the IR. P1 might also contain a carbonyl group, but that was never determined, and neither was the number of exchangeable protons in P2, which could be two. No matter how one attempts to use the above-described classification system, one cannot express this information accurately.

In the new approach, for which the algorithmic design has been completed, one is allowed to express data in a much more natural sequence which parallels the experimental steps. The first experimental step after a reaction is usually the separation and purification of products. An analogous step is to be included in REACT, in which the separation amounts to the setting up of a specified number of labelled "flasks" (analogous to the labels P1 and P2 in the above example) each of which is ultimately to contain a specified number (usually 1) of the products. As experimental data are gathered on each real product, corresponding substructure constraints are attached to the corresponding flask in the program. As each such assertion is made, the bookkeeping mechanism verifies that, for a set of reaction products from a given starting material, there is at

least one way of distributing them among the flasks such that each product satisfies the constraints for its flask. If this test is ever violated, the starting material is removed as a candidate structure. Flasks containing more than one product may be further separated into "subflasks" to any level, and the contents of any flask may be made to undergo further reactions. This capability, the reacting of flask contents, is analogous to common laboratory procedures in which incomplete separations of products are encountered. Dealing with such situations adds considerable complexity to the bookkeeping mechanism, because the contents of a flask may be ambiguous to the program when the reaction is applied. REACT must keep track of all possible structures which might, based on the current flask constraints, occupy the reacting flask. If such a reaction fails (because the products did not satisfy the constraints specified for them), REACT does not eliminate the starting structure entirely, but notes that the structure may not occupy that flask in future flask-allocation tests.

#### 4.2.2 MSRANK

This program is an outgrowth of MSPRUNE described in last year's annual report. It is a combination of a predictor which uses a very simple theory of mass spectrometry to predict the spectra of candidate structures, and an evaluation function which compares the predictions with the observed spectrum of the unknown, assigning a goodness-of-fit score to each candidate. The candidates are then sorted based upon how well they match the observations. The basic concept here is not a new one to the DENDRAL project [see, for example, Buchanan, et al. in Machine Intelligence 4 (Meltzer & Michie, eds., Edinburgh Univ. Press, 1969)], but there are some new aspects to the problem when viewed in the overall CONGEN context.

Because of the wide variety of structural types which can be produced by CONGEN, it is necessary for MSRANK to use a very general model of mass spectrometry. The best predictive theories of mass spectrometry are limited to families of closely related structures (i.e., class specific theories), and the Meta-DENDRAL program is designed to help in discovering such theories. There are very few general principles upon which to draw in predicting mass spectra, though, so MSRANK is limited to only the most approximate kinds of evaluation functions. One principle which we noticed being used by practicing mass spectrometrists was: of two candidate structures for an unknown, the most likely structure is the one which explains the observations most "simply" - i.e., with the fewest complex explanations involving many bond cleavages and the transfer of many hydrogen atoms. The evaluation function used by MSRANK is based on a quantitation of this principle.

In predicting a spectrum, MSRANK explores all possible cleavages of the molecule within some very general user-defined constraints concerning the number of bonds broken and the number of steps in a process, the proximity of pairs of cleaved bonds (i.e., whether or not two adjacent bonds can break in a given process) and the multiplicity or aromaticity of each cleaved bond. Within these general limits, the user also supplies numerical plausibilities from 0 to 1 on the various kinds of breaks which are allowed to occur. For example, he might give unit plausibility to 1-bond cleavages, .8 to 2-bond processes and .6 to 3-bond processes. Aromatic-bond, multiple-bond and adjacent-bond cleavages, if allowed, are given separate plausibilities, as are the allowed neutral transfers. MSRANK combines these values multiplicatively in evaluating the overall plausibility of a specific mass spectral process, and that value is associated with the corresponding predicted mass point. If two different processes predict the same mass point, the higher plausibility value is retained. The result is a predicted spectrum with numbers attached to each peak, interpreted roughly as the "reasonableness" or "simplicity of prediction" measure.

We expect such a theory to be overly complete in the sense that, when applied to the correct structure for an unknown, it will doubtless predict many plausible peaks for which there is no observation. This simply reflects the fact that the "break everything" approach to mass spectrometry is a considerable oversimplification. Thus the evaluation function does not penalize for predicted but unobserved peaks. What we do expect, though, is that a large number of the observed peaks, particularly the intense ones, will have plausible explanations with respect to the correct structure. Thus a "reward" is given to every observed peak which is predicted, the amount being proportional to the plausibility of the prediction and (at the user's option) to the intensity and/or mass value of the observed peak. The sum of rewards for all observed peaks then constitutes the overall score for the candidate which gave rise to the predicted spectrum.

MSRANK is quite new and we have not yet had sufficient experience with it to evaluate its overall usefulness. By using only unit plausibilities for selected characteristics of the mass-spectral cleavages, we are able to duplicate earlier results obtained with the predictor/comparator functions applied to mono- and di-ketoandrostanes. These tests serve to check the accuracy of the MSRANK program. We are now doing a systematic study of various classes of compounds by ranking the spectrum of a known structure against a CONGEN-generated list of structures which contains the correct one among several which are closely related.

## 5 USE OF CONGEN BY OTHER SCIENTISTS

The number of persons experimenting with CONGEN has grown as a result of both the continuing practice of issuing an "invitation for program trial use" at the conclusion of publications, as well as continuing personal contact between Dendral project members and potential program users. Three categories of users make up this group:

### 5.1 Chemists Using Exported Programs

The part of CONGEN responsible for teletype output of chemical structures (the DRAW program) is coded in Fortran. Since the paper describing this program appeared in print [R. Carhart, JACS, 16:82, 1976], we have exported the program to half a dozen sites, ranging from Japan, across North America, to England. Similarly, the entire CONGEN program, is largely coded in Interlisp and SAIL, and has been exported to a collaborator in England who is very interested in the methods and programming techniques employed in coding the program. Another program which we have exported for use by other chemists is the PDP-11 CLEANUP program which was described in ANALYTICAL CHEMISTRY [48:1368, 1976]. This program "cleans up" new GC/MS data to eliminate noise peaks and to separate the data associated with components in the mixture.

In each case, the requestors were provided with an initial choice of format options from which they could select the one most suitable for their computer installation. They were asked to send a 2400 foot reel of magnetic tape appropriate to the selected format option. The programs were written on the tape and returned to them along with a brief written explanation of program organization. Accurate records are kept of who has received the programs, so that omissions and errors can be corrected by mail at a later date, if ever necessary.

1. Dr. James F. Elder, Dow Chemical U.S.A., Midland, Michigan.
2. Dr. Robert M. Supnik, Massachusetts Computer Associates, Inc., Wakefield, Massachusetts.
3. Mr. Dan Pearce, Orange County Sheriff-Coroner Department, Santa Ana, California 92702
4. Dr. H. J. Stoklosa, Central Research & Development Department, E. I. du Pont de Nemours & Company, Wilmington, Delaware.
5. Dr. Douglas W. Kuehl, Environmental Research Laboratory-Duluth, Duluth, Minnesota.



6. Dr. Richard A. Graham, Food Sciences Laboratory, U. S. Army Natick Laboratories, Natick, Massachusetts.
7. Dr. Walter M. Shackelford, United States Environmental Protection Agency, Environmental Research Laboratory, Athens, Georgia.
8. Dr. Richard Gans, Chemical Research Division, American Cyanamid Company, Bound Brook, New Jersey.
9. Dr. John C. Marshall, Department of Chemistry, the University of North Carolina, Chapel Hill, North Carolina.
10. Dr. Graham S. King, Department of Chemical Pathology, Queen Charlotte's Hospital for Women, London, England.
11. Dr. J. Wyatt, Chemistry Division, Naval Research Laboratory, Washington, D. C..
12. Dr. Gareth Templeman, Research and Development Laboratories, The Pillsbury Company, Minneapolis, Minnesota.
13. Dr. J. B. Justice, Department of Chemistry, Emory University, Atlanta, Georgia.
14. Dr. Thomas Knudsen, Northrop Services, Environmental Sciences Group, Research Triangle Park, North Carolina.
15. Dr. Ingolf Meineke, Fachbereich Chemie, Philipps Universitaet, Lahnberge, West Germany.
16. Dr. M.A. Shaw, Unilever Research, Port Sunlight Laboratory, Wirral, Merseyside, England.
17. Dr. Ernst Weber, Varian MAT, Bremen, West Germany.
18. Paul V. Fennessey, Department of Pediatrics, University of Colorado Medical Center, Denver, Colorado.
19. R. G. A. R. MacLagan, Department of Chemistry, University of Canterbury, Christchurch, New Zealand.
20. James E. Oberholtzer, Arthur D. Little, Inc., Cambridge, Massachusetts.
21. F. Street, AEI Scientific Apparatus Limited, Manchester, England.

## 5.2 Remote Users of SUMEX

Due to the fact that the SUMEX computer is available via both the TYMNET and ARPANET communication networks, it is possible for scientists in many parts of the world to directly access the Dendral programs on SUMEX. Primary usage is centered on CONGEN, although INTSUM is beginning also to gain a following. Although access points to SUMEX are widespread, they frequently are not diverse enough to accommodate the dispersed group of scientists who have expressed an interest in using one of the Dendral programs. For example, Dr. Joseph Baker of the Roche Institute of Marine Pharmacology in Dee Why, Australia, is looking at the possibility of accessing SUMEX by using International Direct Distance Dialing (IDDD).

## 5.3 Chemists Communicating by Mail

Many Scientists interested in using DENDRAL programs in their own work are not located near a network access point. Users of this type choose to use the mail to send details of their structure elucidation problem to a Dendral Project collaborator at Stanford.

## 5.4 Chemical Problems Posed to CONGEN

Following is a list of CONGEN users, and a brief summary of their program interests during the past year.

1. Dr. Roger Hahn, Syracuse University. While at Stanford he used CONGEN to help solve the structures of photoproducts by obtaining all possibilities under available constraints and designing NMR experiments to differentiate the possibilities. This work will be published soon.
2. Dr. William Epstein, University of Utah. During a demonstration of CONGEN, he posed a problem to verify that the structural possibilities he determined for an unknown were in fact all possibilities. The structure of methyl santolate has been published (see Epstein, et al., J.C.S. Chem. Commun., 590 (1975)).
3. Dr. Clair Cheer, University of Rhode Island. While on sabbatical at Stanford, Dr. Cheer has worked on a number of structure elucidation problems using CONGEN including Briareine D and [+]-Palustrol (Cheer et al., Tetrahedron Letters, 1807 (1976)). Work is